

Neural EEG-Speech Models

Pengfei Sun, and Jun Qin,

Abstract

In this paper, we describe three neural network (NN) based EEG-Speech (NES) models that map the unspoken EEG signals to the corresponding phonemes. Instead of using conventional feature extraction techniques, the proposed NES models rely on graphic learning to project both EEG and speech signals into deep representation feature spaces. This NN based linear projection helps to realize multimodal data fusion (i.e., EEG and acoustic signals). It is convenient to construct the mapping between unspoken EEG signals and phonemes. Specifically, among three NES models, two augmented models (i.e., IANES-B and IANES-G) include spoken EEG signals as either bias or gate information to strengthen the feature learning and translation of unspoken EEG signals. A combined unsupervised and supervised training is implemented stepwise to learn the mapping for all three NES models. To enhance the computational performance, three way factored NN training technique is applied to IANES-G model. Unlike many existing methods, our augmented NES models incorporate spoken-EEG signals that can efficiently suppress the artifacts in unspoken-EEG signals. Experimental results reveal that all three proposed NES models outperform the baseline SVM method, whereas IANES-G demonstrates the best performance on speech recovery and classification task comparatively.

Index Terms

EEG-Speech mapping, multimodal, unsupervised, supervised, gated neural networks.

I. INTRODUCTION

Human-computer-interfaces (HCIs) involve electroencephalography (EEG) based communication [1]. By translating EEG patterns (i.e., user intent) into control signals, HCIs are used to operate machines such as computers or assistive devices. As one of major brain activities, imagined speech can be a promising carrier to enforce intuitive HCIs. Recently, automatic speech recognition technique has facilitated our lives by providing a new and more natural means of communication with machines. Previous research [2] demonstrates that it is feasible to use EEG for the recognition of human speech, and broaden the boundary of human-machine perception.

The key issue of EEG based speech recognition is to model the relationship between EEG signals and speech parameters. Despite various studies [3], [4] demonstrate that speech activities can be represented by EEG data, it is still very difficult to correlate the two types of signals. The major challenges [5] include: 1) EEG signals are easily contaminated by artifacts produced by motion and muscle movements appearing from spoken speech, and 2) the relevant brain activities (e.g., prior or posterior emotions) significantly affect the presenting forms of speech related EEG signals. Therefore, the observed EEG signals generally are considered as the mixture of multiple sources, and speech associated EEG signals are masked by other brain activities. The features (e.g., mean, kurtosis and entropy)

used in the conventional EEG-Speech classification techniques, may not yield deep representation of speech related EEG data.

Another practical issue for EEG-Speech models is the multimodal data fusion, in which EEG features should be translated to speech features. One popular approach is to symbolize each acoustic phoneme or word as single label. Such type of feature dimension reduction obviously cannot reflect the physical responses of acoustic signals to the corresponding EEG signals, which as a result may lose the accuracy and resolution provided by EEG scalp distribution. Moreover, in the higher dimensional acoustic data space, it may be much easier to construct the unique mapping relationships between EEG-Speech pairs.

In this study, we propose three neural network (NN) based EEG-Speech (NES) models. In the first NES model, a joint directed-undirected graphical structure is proposed. This NN consists of top-bottom linear transformation layers, which are connected by a restricted Boltzmann machine (RBM). Unspoken-EEG signals and phonemes are the input and output, respectively. The second NES model augments the first model by incorporating the spoken EEG signals to bias the unspoken EEG signals at the input layer. Most previous EEG-Speech models treat the spoken and unspoken EEG signals separately when constructing the translation model. The third one is based on the second biased model, in which a three-way gated NES model is described. In this multiplicative model, the spoken EEG signals are used not only to bias but also to condition the unspoken EEG signals. By utilizing the multiplicative modulation, the third model can suppress the uncorrelated background EEG signals. A factored RBM training approach [6] is also introduced to enforce the learning process.

The rest of the paper is organized as follows: the related works will be presented in Section II. In section III, we will show how to learn EEG representations and speech features together by training our EEG-Speech models in a joint unsupervised and supervised manner. Experimental details and relative evaluation will be discussed in section IV.

II. RELATED WORK

The related work can be cataloged into two groups: 1) EEG signals based speech classification and recognition, and 2) multimodal representation learning.

Comprehensive studies have been conducted in EEG based speech classification and recognition. In the branch of imagined speech recognition, various voice related potentials, i.e., vowels, phonemes, syllables, and whole-word [7]–[10], are used as speech labels. Porbadnigk *et al.* [11] applied an HMM to classify EEG signals associated with consecutive words, and the limited performance revealed that the order of words significantly affected the results. In other studies, unspoken EEG signals have also been used to decode variety of languages [8], [12], [13]. Recently, Zhao and Rudzicz [14] investigated spoken EEG signals based speech classification by utilizing a deep belief network. In their model, the inputs were still shallow features gathered by conventional techniques. Di Liberto *et al.* [2] and O’Sullivan *et al.* [15] both worked on spoken EEG translation and devoted on directly mapping to the acoustic features of speech (i.e., envelope or spectrogram). Yoshimura *et al.* incorporated fMRI data as a hierarchical prior for EEG feature extraction by using a variational Bayesian method [16]. Their results showed

that adding extra imagined speech information can boost the EEG based recognition. In this study, the similar idea has been incorporated into our NES models by combining both spoken and unspoken EEG signals.

With respect to multimodal representation learning, several deep learning methods have been introduced to successfully extract the intrinsic structures from multiple modalities. Ngiam *et al.* [17] learned features from audio and video by utilizing deep autoencoders. Srivastava *et al.* [18] developed a model of images and text relying on the multimodal deep Boltzmann machine. More recently, Socher *et al.* [19] and Frome *et al.* [20] proposed methods for mapping images into a text representation space. The idea of domain transformation in their work quite conforms to our approach that we project both EEG and speech signals into the corresponding feature spaces. Specifically, the feature space of EEG signals is learned from a context model derived from the concept of log-bilinear language model [21]. In addition, Kiros *et al.* [22] proposed a 3-way network structure to enforce the joint image-text learning. By adding extra gate branch, such network structure can efficiently incorporate the underlying correlations. In our proposed models, the gated NN are introduced to incorporate the relationships between unspoken and spoken EEG signals.

III. EEG-SPEECH MODELS

A. Imagined Neural EEG-Speech Model (INES)

A typical imagined EEG-Speech model is mapping the multichannel EEG signals to the speech label or the relevant acoustic features. Compared with conventional classification or transformation approaches, in our proposed INES model, each EEG channel is treated as the context of its neighbor channel as shown in Fig. 1. As a result, the EEG signals at the batch t_k are considered as a tuple, that is, the $n-1$ channel EEG signals w_1, \dots, w_{n-1} are followed by corresponding speech component w_n . It is based on the hypothesis that a unique unspoken EEG pattern corresponding to each individual speech segment exists, and is bundled with a series of EEG signals.

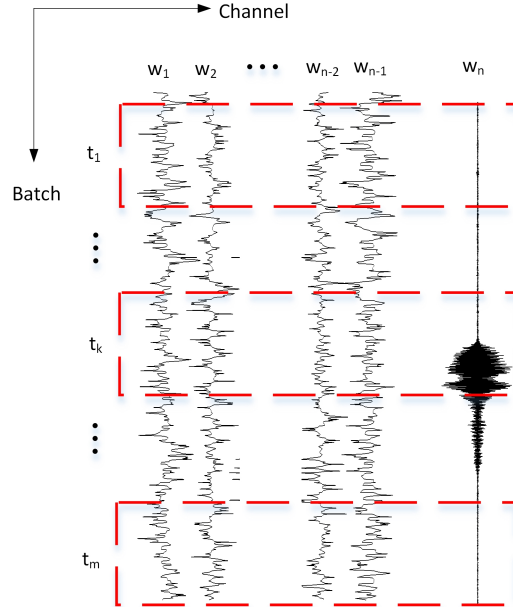


Fig. 1. The schematic diagram of joint-context EEG-Speech tuple. The horizontal axle is the channel label and the vertical axle is time batch.

To define the middle layer RBM model for the distribution of the acoustic feature w_n , given the context $w_{1:n-1}$, an energy function must be specified for a joint configuration of the visible and hidden units. To realize the modal fusion between EEG signals and speech components, we project these two types of modals into the same feature space. Therefore, as shown in Fig. 3, before directly feeding the EEG signals into NN, each channel w_i is represented as a D -dimensional real-valued vector \mathbf{x}_w .

$$\hat{\mathbf{x}} = \sum_{i=1}^{n-1} \mathbf{x}_{w_i} \mathbf{F}^{(i)} \quad (1)$$

where $\mathbf{F}^{(i)}$, $i = 1, \dots, n-1$ are $D \times D$ feature parameters matrices. Thus, $\hat{\mathbf{x}}$ is the feature representation of the EEG signals. Accordingly, after obtaining the hidden layer output \mathbf{h} , a linear transformation is enforced to project \mathbf{h} as the acoustic modal features that can be described as $\hat{\mathbf{h}} = \mathbf{h}\mathbf{J}$. The transformation matrix \mathbf{J} is $K \times L$, where K is the dimension of acoustic feature space and L is the dimension of phoneme vector. Briefly, we start by specifying the energy function:

$$E(\mathbf{h}; w_{1:n-1}) = -\hat{\mathbf{x}}\mathbf{W}^T\mathbf{h} - \mathbf{b}_h^T\mathbf{h} - \mathbf{b}_x^T\hat{\mathbf{x}} \quad (2)$$

The vector \mathbf{b}_h contains the biases for the hidden units, while vectors \mathbf{b}_x contain biases for EEG signals. The conditional distribution of the hidden configuration \mathbf{h} is defined in terms of the energy function as

$$P(\mathbf{h}|w_{1:n-1}) = \frac{1}{Z_c} \exp(-E(\mathbf{h}; w_{1:n-1})) \quad (3)$$

where $Z_c = \sum_{\mathbf{h}} \exp(-E(\mathbf{h}; w_{1:n-1}))$ is a context-dependent normalization term. In INES model (as shown in Fig.3 (a)), the feed-forward NN structure is mixed with an undirected RBM, which is applied for representing this mapping model. Unlike the existing EEG based speech classification techniques, the proposed INES model learns the deep feature representations relying on its top-bottom NN structure. In the context of a feed-forward NN, the weights between the EEG features $\hat{\mathbf{x}}$ and \mathbf{h} is the representation transformation matrix \mathbf{W} . The output \mathbf{h} then can be inversely projected as phoneme. Learning process is a combination of standard back propagation based supervised training and contrast divergence based unsupervised training.

B. Imagined-Articulated Neural EEG-Speech Models

Supposing that along with each training tuple of imagined speech related EEG signals (w_1, \dots, w_{n-1}) , there is an associated vector $\mathbf{y} \in \mathcal{R}^M$ corresponding to the feature representation of the modality to be conditioned on, such as an articulated-speech related EEG. By utilizing the spoken EEG signals to strengthen the feature extraction of unspoken EEG signals in the INES model, we propose two different Imagined-Articulated Neural EEG-Speech (IANES) models. In following subsection, two different IANES models (as shown in Fig. 3 (b) and (c)) will be described, respectively.

1) *Biased Imagined-Articulated Neural EEG-Speech Model (IANES-B)*: The IANES-B model is a straightforward extension of the INES model. The IANES-B model can be viewed as a feedforward network by taking the INES network and adding a context channel based on articulated spoken EEG signals related features \mathbf{y} . Because the

artifacts existed in unspoken EEG signals can easily mislead to wrong translation of the EEG-Speech model, the introduced spoken-EEG signals can bias the deviation caused by such artifacts.

The IANES-B model introduces an additive bias to the recovery acoustic feature representation $\hat{\mathbf{h}}$, and can be described as:

$$\hat{\mathbf{x}} = \left(\sum_{i=1}^{n-1} \mathbf{x}_{w_i} \mathbf{F}^{(i)} \right) + \mathbf{y} \mathbf{M}^m \quad (4)$$

where \mathbf{M}^m is a $M \times D$ feature matrix that projects the spoken EEG signals into the feature space \mathcal{W}^D . M is the product of $n - 1$ and D . This model has similar training procedure as INES model: given acoustic feature representation $\hat{\mathbf{x}}$, computing the conditional distribution $P(\mathbf{h}|w_{1:n-1}, \mathbf{y})$ of \mathbf{h} remains the same. The bias branch can be updated when back propagation training is implemented.

2) *Gated Imagined-Articulated Neural EEG-Speech Model (IANES-G)*: As a more powerful model to incorporate articulated-speech related spoken EEG conditioning, the proposed IANES-G model is to gate the representation matrix \mathbf{W} by the features \mathbf{y} . By doing this, \mathbf{W} becomes a tensor. Therefore, \mathbf{W} requires $D \times M \times K$ parameters, which is impractical for general network training. A solution is to factor \mathbf{W} into three lower-rank matrices \mathbf{W}_{fx} , \mathbf{W}_{fy} and \mathbf{W}_{fh} . For each acoustic feature \mathbf{h} can specify its own hidden and gated weight matrix. More specifically, $(\mathbf{W}_{fx})^T \mathbf{W}_{fy}$ denote the EEG embeddings. Given a EEG feature representation $\hat{\mathbf{x}}$, the factor outputs are

$$f = (\mathbf{W}_{fx} \hat{\mathbf{x}}) \cdot (\mathbf{W}_{fy} \mathbf{y}) \quad (5)$$

where \cdot is a element-wise product. Looking into (5) helps to understand the advantage of IANES-G model: the multiplication is beneficial to strengthen the highly correlated components between unspoken EEG and spoken EEG signals, and as a result it can greatly suppress those EEG features with less similarity. Comparing with IANES-B model, IANES-G model is more efficiency to catch the underlying features of unspoken and spoken EEG signals. However, comparing with additive biased model, one weakness for IANES-G model is its sensitivity to local faked features. For instance, the spikes related to non-speech brain activities may be easily classified as speech associated EEG signals when applying IANES-G model.

C. Gated restricted Boltzmann machine learning

The three way tensor \mathbf{W} in the IANES-G model can be factored into decoupled matrices, and accordingly the energy function can be given as [6]

$$\begin{aligned} -E(\mathbf{y}, \mathbf{h}; \mathbf{x}) &= \sum_k b_k^h h_k + \sum_j \frac{(y_j - b_j^y)^2}{2\sigma_j^2} + \sum_i \frac{(x_i - b_i^x)^2}{2\sigma_i^2} \\ &+ \sum_f \left(w_{if}^x \sum_i \frac{x_i}{\sigma_i} \right) \left(\sum_j w_{jf}^y \frac{y_j}{\sigma_j} \right) \left(\sum_k w_{kf}^h h_k \right) \end{aligned} \quad (6)$$

By noting

$$f_f^x = \sum_{i=1}^D w_{if}^x \frac{x_i}{\sigma_i}, \quad f_f^y = \sum_{j=1}^M w_{jf}^y \frac{y_j}{\sigma_j}, \quad f_f^h = \sum_{k=1}^K w_{kf}^h h_k \quad (7)$$

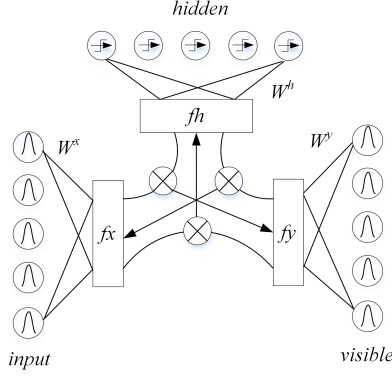


Fig. 2. The schematic of symmetrical three-way RBMs. The three factor layers have the same size, and \otimes refers to element-wise multiplication.

where i, j and k index input, visible and hidden units, respectively. The bold font represents the variable, and small cap denotes the observation. x_i and y_j are Gaussian units, and h_k is the binary state of the hidden unit k . σ_i and σ_j are the standard deviations associated with x_i and y_j , respectively. The terms b_k^h and b_j^y represent biases of the hidden and observable units, respectively. Inferences of the k th hidden and j th visible unit can be performed as

$$p(h_k = 1 | \mathbf{y}; \mathbf{x}) = S(\Delta E_k) \quad (8)$$

$$p(y_j = y | \mathbf{h}; \mathbf{x}) = N(y | \Delta E_j, \sigma_j^2) \quad (9)$$

where $N(\cdot | \mu, \sigma^2)$ denotes the Gaussian probability density function with mean μ and standard deviation σ . $S(\cdot)$ is the sigmoid activation function. ΔE_k and ΔE_j are the overall inputs of the k th hidden unit and j th visible unit, respectively [23]. The gated RBMs allows the hidden units to model the transition between successive frames, and generally the input units are collected directly from previous frames in many applications [24].

The three factor layers as shown in Fig. 2 have the same size F , and the factor terms (i.e., $W^x \mathbf{x}$, $W^y \mathbf{x}$, and $W^h \mathbf{h}$) correspond to three linear filters applied to the input, visible, and the hidden unit, respectively. To perform k -step Gibbs sampling in the factored model, the overall inputs of each unit in the three layers are calculated as

$$\Delta E_k = \sum_f w_{kf}^h \sum_i w_{if}^x \frac{x_i}{\sigma_i} \sum_j w_{jf}^y \frac{y_j}{\sigma_j} + b_k^h \quad (10)$$

$$\Delta E_j = \sum_f w_{jf}^y \sum_i w_{if}^x \frac{x_i}{\sigma_i} \sum_k w_{kf}^h h_k + b_j^y \quad (11)$$

$$\Delta E_i = \sum_f w_{if}^x \sum_j w_{jf}^y \frac{y_j}{\sigma_j} \sum_k w_{kf}^h h_k + b_i^x \quad (12)$$

In (10)-(12), the factor layers are multiplied element-wise (as the \otimes illustrated in Fig. 2) through the same index f . These are then substituted in (8)-(9) for determining the probability distributions for each of the visible and hidden units.

Therefore, each speech pattern in the hidden units corresponds to a pairwise matching of input filter responses and visible filter responses. The learning procedure aims to find a set of filters that can reflect the correlations of consecutive speech frames in the training data.

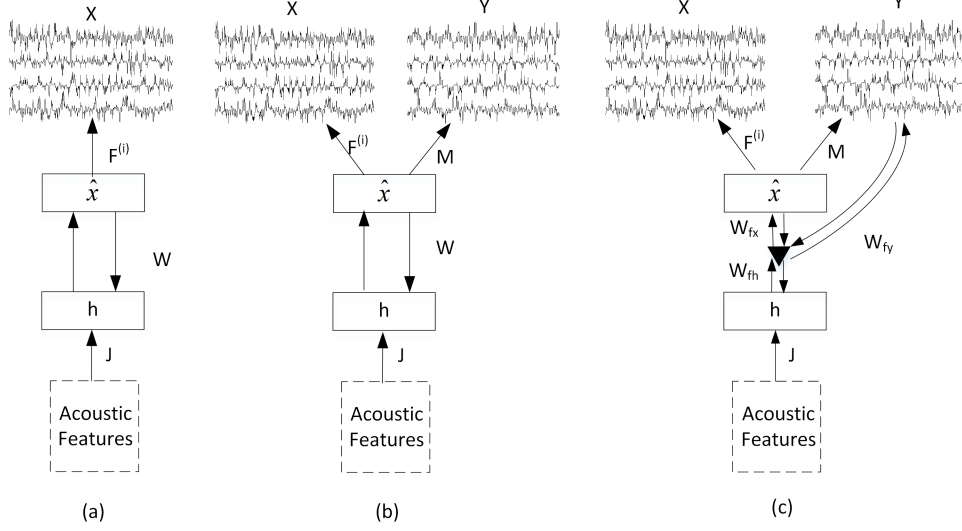


Fig. 3. The schematic diagram of three proposed neural network based EEG-Speech models, (a)INES, (b)IANES-B, and (c)IANES-G, respectively. The arrow in the figure represents the training flow.

To train factored RBMs, one needs to maximize the average log-probability $L = \log p(\mathbf{y}|\mathbf{x})$ of a set of training pairs $\{(\mathbf{x}, \mathbf{y})\}$. The derivative of the negative log-probability with respect to parameters θ is given as

$$-\frac{\partial L}{\partial \theta} = \left\langle \frac{\partial E(\mathbf{y}, \mathbf{h}; \mathbf{x})}{\partial \theta} \right\rangle_{\mathbf{h}} - \left\langle \frac{\partial E(\mathbf{y}, \mathbf{h}; \mathbf{x})}{\partial \theta} \right\rangle_{\mathbf{h}, \mathbf{y}} \quad (13)$$

where $\langle \cdot \rangle_{\mathbf{v}}$ denotes the average with respect to variable \mathbf{v} . In practical, Markov chain step running is used to approximate the averages in Eq. (13). By differentiating (6) with respect to the parameters, we get

$$-\frac{\partial E}{\partial w_{kf}^h} = -h_k \sum_i x_i w_{if}^x \sum_j y_j w_{jf}^y \quad (14)$$

$$-\frac{\partial E}{\partial w_{jf}^y} = -y_j \sum_i x_i w_{if}^x \sum_k h_k w_{kf}^h \quad (15)$$

$$-\frac{\partial E}{\partial w_{if}^x} = -x_i \sum_j y_j w_{jf}^y \sum_k h_k w_{kf}^h \quad (16)$$

$$-\frac{\partial E}{\partial b_k^h} = h_k, \quad -\frac{\partial E}{\partial b_i^x} = x_i, \quad -\frac{\partial E}{\partial b_j^y} = y_j \quad (17)$$

In this study, we use the power of EEG and speech signals as the input \mathbf{x} and output $\hat{\mathbf{h}}$, and to encourage nonnegativity in three factored matrices w_{kf} , w_{if} , and w_{jf} , a quadratic barrier function is incorporated to modify the log probability, that is, the objective function is now the following regularized likelihood [25]

$$-\mathcal{L}_{reg} = \mathcal{L}(\mathbf{y}; \mathbf{x}) - \frac{\alpha}{2} \sum \sum f(w) \quad (18)$$

where

$$f(x) = \begin{cases} x^2 & x < 0 \\ 0 & x \geq 0 \end{cases}$$

A joint unsupervised and supervised training is used to map unspoken-EEG signals \mathbf{x} to phonemes $\hat{\mathbf{h}}$ stepwise. Specifically, the linear feature transformations as parts of the NN are also updated during the network learning. For all three NES models, the training sequence to obtain the network parameters can be given in the following algorithm framework

Algorithm 1: Joint unsupervised-supervised training

```

for  $iteration \leq N_{epoch}$  do
  for  $Iteration \leq N_{batch}$  do
    Forward
     $\hat{\mathbf{x}} \leftarrow \mathbf{x}_{batch}$  ;
     $\mathbf{y}_{bias} \leftarrow \mathbf{y}$ ;
     $\hat{\mathbf{x}} = \hat{\mathbf{x}} + \mathbf{y}_{bias}$ ;
    Unsupervised
    sample  $\mathbf{h} \sim p(\mathbf{h}|\mathbf{y}^t, \mathbf{x}^t)$  by (8)(10) ;
    calculate  $\langle \frac{\partial E}{\partial \theta} \rangle_{\mathbf{h}}$  by (14)-(17);
    for  $iteration \leq N_{step}$  do
      sample  $\mathbf{h}^{t,n} \sim p(\mathbf{h}|\mathbf{y}^{t,n}, \mathbf{x}^{t,n})$  by (8)(10);
      sample  $\mathbf{y}^{t,n} \sim p(\mathbf{y}|\mathbf{x}^{t,n}, \mathbf{h}^{t,n})$  by (9)(11);
      calculate  $\langle \frac{\partial E}{\partial \theta} \rangle_{\mathbf{h}, \mathbf{y}}$  by (14)-(17) ;
      update  $\{w_{fu}, w_{fs}, w_{fa}, \mathbf{b}_h, \mathbf{b}_x\}$  by (13) and (18);
       $\hat{\mathbf{h}} \leftarrow \mathbf{h}$  ;
    Backward
     $\Delta = \|\hat{\mathbf{h}} - target\|$  ;
     $\frac{\partial \Delta}{\partial J}, \frac{\partial \Delta}{\partial W}, \frac{\partial \Delta}{\partial M}, \frac{\partial \Delta}{\partial F^{(i)}}$ ;
    update  $\{F^{(i)}, M, J, W\}$  ;

```

Output: parameter set

IV. EXPERIMENTS

In this study, we perform experimental evaluation of the proposed three NES models on a public available dataset KARA ONE [14]. This dataset combines 3 modalities (EEG, face tracking, and audio) during imagined and vocalized phonemic and single-word prompts. In their EEG recording experiments, each participant was instructed to look at the computer monitor and move as little as possible. Seven phonemic/syllabic prompts and 4 words were used in several repeated experimental trails. Each trail consisted of 4 successive states, including a rest state, a stimulus state, an imagined speech state, and a speaking state. The EEG data were collected by using a 64-channel

Neuroscan Quick-cap, and the electrode placement followed the 10-20 system. In this study, two channel (i.e., O1 and O2) are not used and rest of 62 channel EEG data are used for evaluations of the proposed NES models.

A. Experimental Data Process

A general data pre-process is applied to the EEG data, including removal of ocular artifacts using blind source separation [26]. The EEG data are band-pass filtered between 1 Hz and 50 Hz, and the mean values are subtracted from each channel data. The EEG data are segmented into different trials, and each trial is further segmented into the 4 states described in above section, among which the EEG signals associated with imagined speech and articulated speech are used as the feed. All EEG data are normalized in each channel, and in each channel the window size is fixed to 20 ms and a hop size 10 ms (i.e., 100 segments in each second). The speech signals are downsampled into 1kHz to match the sampling rate of EEG signals.. After the pre-filtering, the envelop of speech is extracted as the training data. To reduce the time shift between speech components and EEG signals, the two types of signals are padded with the same length.

B. Training process

All proposed NES models are trained using the following hyperparameters: all 'context' channel matrices use a weight decay of 1.0×10^{-4} while EEG representations use a weight decay of 1.0×10^{-5} . All other weight matrices, including the GRBM filters, use a weight decay of 1.0×10^{-4} . In each EEG state, the batch size is set as 2000 and an initial learning rate is 0.1. Our IANES-G model uses an initial learning rate of 0.02. Initial momentum is set to 0.5 and is increased linearly to 0.9 over 30 epochs. The EEG representation matrices are initialized to the 50 dimensional features. The 62 channel EEG data are all input once as the co-context for each other.

In each experiment, we split the dataset into 80% as training set and 20% as validation set. The NES models are learned by a joint supervised and unsupervised training process, in which the middle layer (i.e., GRBM) is unsupervised initialized, and then the parameter sets are updated by supervised backward training. Specifically, the two procedures are implemented stepwise. The linear transformation F , M and J are all updated based on the NN parameter back propagation.

C. Recovery of Phoneme

The main goal of this study is to map unspoken EEG signals with speech signals. Figure 4 shows the recoveries of four different phonemes, including two typical phonemic prompts ($/uw/$, $/iy/$) and two phonetically-similar pairs (pat , and $knew$), from 62 channels EEG signals recorded on a single participant (MM05). The output signal \hat{h} is regarded as the recovery of the speech envelop, and both speech signals and recovery output \hat{h} are normalized.

As shown in Fig. 4, all three proposed NES models can reflect the major fluctuations of the speech components. Specifically, the IANES-B and IANES-G models demonstrate similar performances, and both catch more details

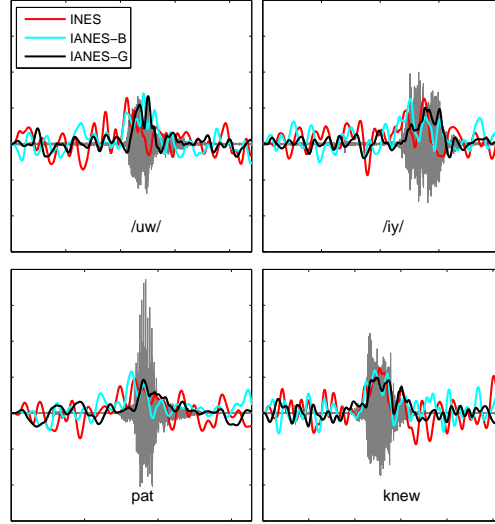


Fig. 4. The original speech components (gray) and the recovery speech envelopes by three proposed NES models. Four different phonemes, including */uw/*, */iy/*, *pat*, and *knew*, are used, and all 62 channels EEG data recorded on a single participant (MM05) are used for training the NES models.

of speech transitions than the INES model. In addition, the IANES-G model shows the lowest degree of deviation compared with both INES and IANES-B models.

Figure 5 shows the recoveries of four different phonemes, */uw/*, */iy/*, *pat*, and *knew*, from ten EEG signals recorded from 10 selected channels (i.e., FC6, FT8, C5, CP3, P3, T7, CP5, C3, CP1, and C4), which have higher correlations with speech activities. This selected channels experiment confirm the involvement of selected regions during the planning of speech articulation [14].

As shown in Fig. 5, all three proposed NES models obtain significant improvements on recoveries of the speech components. The reason is that, the less-correlated EEG signals have been excluded from the computations. For */iy/*, all three proposed NES models recover the speech component very well. For recovering */uw/*, *pat* and *knew*, the IANES-G model is slightly better than the IANES-B model, while both models gain advantages against the INES model with respect to the envelop recovering and time shifting. The results indicate that the gated network structure provides a stronger constraint to eliminate the useless information.

D. Classification Results

To achieve accurate EEG based speech recognition, correctly classifying EEG pattern into each categories of speech components is the key issue. Most previous studies focused on binary classification due to the fact that it is hard to construct the unique relationships between EEG and speech signals. In this study, we attempt to extend the EEG based speech classifications into a multi-categories scenario.

In this section, we evaluate the proposed IANES-G model for speech classifications by using 50 samples for each single speech label. These speech samples are randomly selected from 14 participants. In order to avoid samples

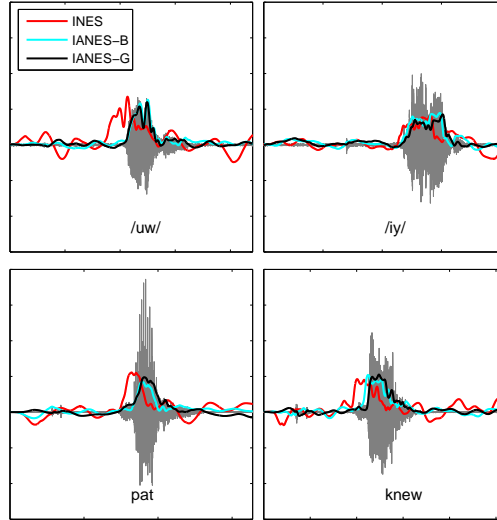


Fig. 5. The original speech components and the recovery speech envelopes by three proposed NES models. Four different phonemes, including */uw/*, */iy/*, *pat*, and *knew*, are used, and 10 selected EEG channels data recorded on a single participant (MM05) are used for training the NES models.

Predict	/uw/	28	1	5	0	3	3	2	2	4	2	2	53.8%
	/tiy/	1	22	2	2	8	2	2	3	5	1	6	40.7%
	/iy/	0	3	21	0	4	4	1	4	3	3	3	45.7%
	/m/	5	2	2	24	2	1	5	2	1	2	4	48.0%
	/n/	3	1	5	4	20	2	1	1	3	3	5	41.7%
	/piy/	1	6	5	1	6	19	2	5	4	1	0	38.0%
	/diy/	5	5	3	0	4	1	18	4	3	2	7	34.6%
	gnaw	4	6	0	1	4	4	4	21	3	2	3	40.4%
	pat	4	3	3	2	4	2	3	6	16	4	5	30.8%
	pot	7	6	0	0	1	2	0	3	2	25	4	50.0%
	knew	6	2	2	6	2	1	3	6	3	5	14	28.0%
56.0% 44.0% 42.0% 48.0% 40.0% 38.0% 36.0% 42.0% 32.0% 50.0% 28.0% 41.5%													
/uw/ /tiy/ /iy/ /m/ /n/ /piy/ /diy/ gnaw pat pot knew													
Actual													

Fig. 6. The confusion matrix of eleven phonemes extracted from 14 participants in KARA dataset. For each phoneme, 50 samples are randomly selected from the trials and then fed into the cross-trained IANES-G model. The output is the predicted categories, when comparing with the actual categories.

with certain label from same participant, we set 10 as maximum samples for each participant. The IANES-G model is cross-trained by all EEG-Speech pairs, and each batch data of speech is normalized. Figure. 6 shows the confusion matrix of eleven phonemes. The diagonal cells show the number of correct classifications by the IANES-G model. For example, 28 */uw/* are correctly classified as its original category. Along the horizontal axis, it is the predicted

classes. For instance, out of 52 $/uw/$ predictions, 53.8% are correct. Meanwhile, the actual classes are presented along the vertical axis. For example, out of 50 $/uw/$ cases, 56% are correctly predicted and 44% are predicted as the rest categories. In addition, this confusion matrix reveals that overall 41.5% of the predictions are correct by IANES-G model. The classification results also show that the wrongly classified phonemes are very close to each other. The potential reason may be explained as the selected phonemes are quite similar to each other, for example, $/uw/$ and $/m/$ are with close acoustic form. The proposed model may achieve better performance trained by more diverse EEG-Speech dataset.

TABLE I

THE CLASSIFICATION RESULTS FOR 11 PHONEMES BY FOUR APPROACHES, INCLUDING THREE PROPOSED INES MODELS AND SVM.

	INES	IANES-B	IANES-G	SVM ^{multi}
$/uw/$	0.47	0.53	0.58	0.24
$/tiy/$	0.28	0.37	0.43	0.21
$/iy /$	0.34	0.39	0.41	0.19
$/m/$	0.35	0.31	0.51	0.25
$/n/$	0.27	0.42	0.39	0.21
$/piy/$	0.29	0.31	0.40	0.18
$/diy/$	0.39	0.36	0.39	0.16
gnaw	0.41	0.46	0.45	0.23
pat	0.32	0.40	0.41	0.17
pot	0.35	0.39	0.46	0.27
knew	0.31	0.35	0.33	0.19

Moreover, a comprehensive evaluation is also conducted among three proposed NES models and support vector machine (SVM) baseline in this study. A variant of SVM used for multi-class formulation is tested, given EEG feature vectors x_i and x_j with phoneme labels. This multi-class SVM supports linear kernels only, and the source code can be found at [http : //www.cs.cornell.edu/people/ tj/svmight/svm_multiclass.html](http://www.cs.cornell.edu/people/tj/svmight/svm_multiclass.html). To enhance the performance of SVM, as presented in the work [14], we compute various features over each window, such as the mean, median, standard deviation, variance, maximum, minimum, sum, spectral entropy, energy, skewness, and kurtosis. Moreover, the first and second derivatives are also calculated. The training set includes all the spoken-EEG and unspoken-EEG samples, and for each phoneme in the validation set, 50 samples are randomly selected from the 14 participants.

The classification results have been summarized in Table. I. The IANES-G model achieves the best performance over the two other NES models. It may be explained as that there exists strong correlation between the spoken-EEG signals and unspoken EEG-signals, and when incorporating the spoken-EEG as bias or gate factor can both lead to an significant improvement on unspoken-EEG feature extraction. This strong correlation may also come from the similarity among the phonemes in KARA dataset. The baseline SVM^{multi} shows weak performance since these shallow features cannot reflect the correlations of EEG-Speech pairs. One can easily find that even comparing with

INES model, SVM^{multi} still has a large gap on the classification results.

V. CONCLUSION

In this study, three NN based EEG-Speech models are proposed. A multimodal fusion framework is developed aiming to translate unspoken-EEG signals to the corresponding speech phonemes. The concept of joint-context tuple is introduced to abstract the EEG pattern as a probability distribution. Unlike the conventional feature extraction approaches, the proposed NES models incorporate the representation of EEG signals into the NN structure, in which EEG signals will be first projected into a feature space that is learned by the NN through both supervised and unsupervised training. In the two augmented NES models, referring as IANES-B and IANES-G, the spoken-EEG signals are included to condition unspoken-EEG as a bias channel or gated channel. In addition, the speech phonemes are also projected into the same feature space to realize the modal fusion. The experimental results show that the proposed NES models can recover the speech envelopes well based on the input EEG signals. The EEG based speech classification results indicate that the proposed NES models outperform a baseline method SVM^{multi} on 11 phonological categories. Specifically, IANES-G model averagely obtains the best classification results compared with three other approaches.

REFERENCES

- [1] T. Toda, A. W. Black, and K. Tokuda, "Statistical mapping between articulatory movements and acoustic spectrum using a gaussian mixture model," *Speech Communication*, vol. 50, no. 3, pp. 215–227, 2008.
- [2] G. M. Di Liberto, J. A. O'Sullivan, and E. C. Lalor, "Low-frequency cortical entrainment to speech reflects phoneme-level processing," *Current Biology*, vol. 25, no. 19, pp. 2457–2465, 2015.
- [3] F. Lotte, M. Congedo, A. Lécuyer, F. Lamarche, and B. Arnaldi, "A review of classification algorithms for eeg-based brain–computer interfaces," *Journal of neural engineering*, vol. 4, no. 2, p. R1, 2007.
- [4] S. Gao, Y. Wang, X. Gao, and B. Hong, "Visual and auditory brain–computer interfaces," *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 5, pp. 1436–1447, 2014.
- [5] C. Herff and T. Schultz, "Automatic speech recognition from neural signals: A focused review," *Frontiers in Neuroscience*, vol. 10, 2016.
- [6] P. Sun and J. Qin, "Enhanced factored three-way restricted boltzmann machines for speech detection," *arXiv preprint arXiv:1611.00326*, 2016.
- [7] P. Suppes, Z.-L. Lu, and B. Han, "Brain wave recognition of words," *Proceedings of the National Academy of Sciences*, vol. 94, no. 26, pp. 14 965–14 969, 1997.
- [8] C. S. DaSalla, H. Kambara, M. Sato, and Y. Koike, "Single-trial classification of vowel speech imagery using common spatial patterns," *Neural Networks*, vol. 22, no. 9, pp. 1334–1339, 2009.
- [9] M. DZmura, S. Deng, T. Lappas, S. Thorpe, and R. Srinivasan, "Toward eeg sensing of imagined speech," in *International Conference on Human-Computer Interaction*. Springer, 2009, pp. 40–48.
- [10] M. Lopez-Gordo, E. Fernandez, S. Romero, F. Pelayo, and A. Prieto, "An auditory brain–computer interface evoked by natural speech," *Journal of neural engineering*, vol. 9, no. 3, p. 036013, 2012.
- [11] A. Porbadnigk, M. Wester, and T. S. Jan-p Calliess, "Eeg-based speech recognition impact of temporal effects," 2009.
- [12] L. Hausfeld, F. De Martino, M. Bonte, and E. Formisano, "Pattern analysis of eeg responses to speech and voice: Influence of feature grouping," *Neuroimage*, vol. 59, no. 4, pp. 3641–3651, 2012.
- [13] L. Wang, X. Zhang, X. Zhong, and Y. Zhang, "Analysis and classification of speech imagery eeg for bci," *Biomedical Signal Processing and Control*, vol. 8, no. 6, pp. 901–908, 2013.

- [14] S. Zhao and F. Rudzicz, "Classifying phonological categories in imagined and articulated speech," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 992–996.
- [15] J. A. O’Sullivan, A. J. Power, N. Mesgarani, S. Rajaram, J. J. Foxe, B. G. Shinn-Cunningham, M. Slaney, S. A. Shamma, and E. C. Lalor, "Attentional selection in a cocktail party environment can be decoded from single-trial eeg," *Cerebral Cortex*, vol. 25, no. 7, pp. 1697–1706, 2015.
- [16] N. Yoshimura, A. Nishimoto, A. N. Belkacem, D. Shin, H. Kambara, T. Hanakawa, and Y. Koike, "Decoding of covert vowel articulation using electroencephalography cortical currents," *Frontiers in neuroscience*, vol. 10, 2016.
- [17] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proceedings of the 28th international conference on machine learning (ICML-11)*, 2011, pp. 689–696.
- [18] N. Srivastava and R. R. Salakhutdinov, "Multimodal learning with deep boltzmann machines," in *Advances in neural information processing systems*, 2012, pp. 2222–2230.
- [19] R. Socher, A. Karpathy, Q. V. Le, C. D. Manning, and A. Y. Ng, "Grounded compositional semantics for finding and describing images with sentences," *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 207–218, 2014.
- [20] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, T. Mikolov *et al.*, "Devise: A deep visual-semantic embedding model," in *Advances in neural information processing systems*, 2013, pp. 2121–2129.
- [21] A. Mnih and G. Hinton, "Three new graphical models for statistical language modelling," in *Proceedings of the 24th international conference on Machine learning*. ACM, 2007, pp. 641–648.
- [22] R. Kiros, R. Salakhutdinov, and R. S. Zemel, "Multimodal neural language models," in *ICML*, vol. 14, 2014, pp. 595–603.
- [23] T. Yamashita, M. Tanaka, E. Yoshida, Y. Yamauchi, and H. Fujiyoshi, "To be bernoulli or to be gaussian, for a restricted boltzmann machine," in *ICPR*, 2014, pp. 1520–1525.
- [24] R. Memisevic and G. E. Hinton, "Learning to represent spatial transformations with factored higher-order boltzmann machines," *Neural Computation*, vol. 22, no. 6, pp. 1473–1492, 2010.
- [25] T. D. Nguyen, T. Tran, D. Q. Phung, and S. Venkatesh, "Learning parts-based representations with nonnegative restricted boltzmann machine," in *ACML*, 2013, pp. 133–148.
- [26] G. Gómez-Herrero, W. De Clercq, H. Anwar, O. Kara, K. Egiastian, S. Van Huffel, and W. Van Paesschen, "Automatic removal of ocular artifacts in the eeg without an eeg reference channel," in *Proceedings of the 7th Nordic Signal Processing Symposium-NORSIG 2006*. IEEE, 2006, pp. 130–133.